Automated kinetic feature extraction from Open Access data

Sébastien Siradze, Pedro S. F. Mendes, Laura Pirro, Joris W. Thybaut

Laboratory for Chemical Technology, Ghent University, Ghent, Belgium

Catalyst design plays a crucial role in the development of more sustainable processes and must therefore be as efficient as possible. In the past, catalyst design was mainly carried out via trial and error. This can be improved by the use of kinetic models to understand the chemistry behind the reactions. The traditional approach to kinetic modelling consists of performing experiments, extracting kinetic information from the data based on the researcher's expertise and using that information to construct a kinetic model. Two ways to make the procedure more efficient are, on one hand, using Open Access data shared by other researchers and, on the other hand, automating the step of data interpretation, hence eliminating the need for extensive expertise. The goal of this research is to investigate the state of Open Access catalytic data and to develop a tool which can automatically extract kinetic information from catalytic data.

Many online databases exist where researchers can share and find experimental data. However, no databases focusing on catalytic data were found, so databases targeting all research fields were investigated. It was found that the number of catalytic datasets was still relatively low. Additionally, the datasets have varying structures and often contain errors. Some driving factors towards data sharing suggest that the situation will improve in the future however. For instance, the FAIR principles, an initiative from the research community, aim at making Open Access data easier to find, more structured and more understandable[1].

The optimal kinetic model for a certain dataset reflects the chemical rules related to the reaction and is not simply the best fitting curve. To extract kinetic information from a dataset, features need to be recognized[2] and linked to chemical rules. The first step in automating this procedure is to generate a smooth curve through the data, to mimic how a researcher would draw a smooth curve based on their intuition. This was done using the Python class 'UnivariateSpline' which generates a spline with a certain number of intervals depending on a given smoothing factor (Figures 1.a and 1.b). An algorithm was developed to determine the ideal smoothing factor for the dataset. This algorithm proved to be efficient also in case of 'small' data, which is usually the case for typical datasets in catalysis (Figure 1.c). Afterwards, the first and second derivatives of the curve were calculated and based on their signs, shapes are assigned to the evolution of the data (Figure 1.d). The algorithms were tested on a hydrodeoxygenation dataset as well (Figure 1.e and 1.f).

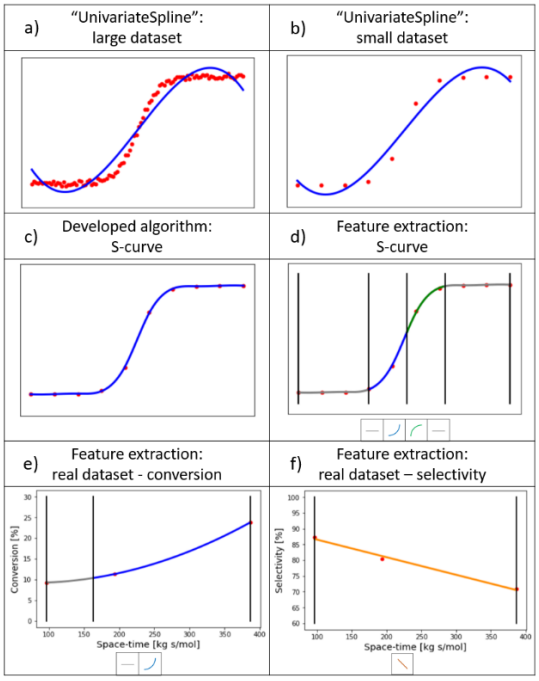

Figure 1: Performance of state-of-the-art and developed algorithms

The identified shapes, together with the values in the extremes of the intervals, represented the desired kinetic features. In a second stage of this work, these will be ranked and used to extract the key information required for the automated construction of a kinetic model.

[1] Wilkinson, M.D., et al., Scientific Data, 3 (2016) 160018.
[2] Janusz, M.E. and Venkatasubramanian, V., Applic. Artif. Intell., 4(5) (1991) 329-339

E-mail : sebastien.siradze@ugent.be